Paraphrase Acquisition from Comparable Medical Corpora of Specialized and Lay Texts

Louise Deléger, MSc[1], Pierre Zweigenbaum, PhD[2]
[1]INSERM, U872, Eq. 20, Paris, F-75006 France; [2]CNRS, LIMSI, Orsay, F-91403 France

Abstract*. Nowadays a large amount of health information is available to the public, but medical language is often difficult for lay people to understand. Developing means to make medical information more comprehensible is therefore a real need. In this regard, a useful resource would be a corpus of specialized and lay paraphrases. To this end, we built comparable corpora of specialized and lay texts on which we applied paraphrasing patterns based on anchors of deverbal noun and verb pairs. The results show that the paraphrases were of good quality (71.4% to 94.2% precision) and that this type of paraphrasing was relevant in the context of studying the differences between specialized and lay language. This study also demonstrates that simple paraphrase acquisition methods can also work on texts with a rather small degree of similarity, once similar text segments are detected.*

Introduction

Nowadays a large amount of health information is available to the general public, be it through the Internet which provides access to medical documents of all kind, or through health records. However, the language used in these documents may not be adapted to people with no or little medical knowledge. Medical language is often difficult for lay people in terms of vocabulary, syntax and structuring of the information. As a result, depending on their level of health literacy, lay people may misunderstand or not understand at all the information they encounter[1,2]. Developing means to make medical information more comprehensible to lay people is therefore a real need[2]. This could mean for instance developing resources to help authoring documents intended for patients or to simplify existing medical texts. In this regard, a useful resource would be a corpus of specialized and lay paraphrases, *i.e.* equivalent lay and specialized expressions. This is the overall goal of the present work.

Empowering lay people with means to understand health information more easily has been the object of several recent studies. A preliminary step to many applications is to examine characteristics of medical texts[3]. Grabar *et al.*[4] used such characteristics to automatically categorize expert vs. non-expert Web pages. McCray *et al.*[5] developed methods to help lay people query the Web for health information. A number of papers emphasized the need for consumer health vocabularies[6,7]. Elhadad[8] identified and defined difficult medical terms. Zeng *et al.*[9] designed a prototype translator to make personal health records comprehensible to patients, simplifying UMLS terms and adding explanations. Building a lexicon of linked specialized and lay expressions was also investigated to bridge the gap between the two language types[10].

In Natural Language Processing several methods have been designed in the field of paraphrasing, most of them dealing with text corpora that differ slightly from ours. Paraphrase acquisition approaches can use parallel corpora, i.e. different translations or versions of the same texts. For instance Barzilay and McKeown[11] used machine translation techniques to detect paraphrases in a corpus of English translations of literary novels. Elhadad and Sutaria[10] also worked with a parallel corpus of medical scientific articles and their lay versions. However such corpora are not easily available and many approaches rely instead on comparable corpora. That is, in the context of paraphrase acquisition, texts from different sources or genres but dealing with the same topic. In this regard, only closely related corpora have been used, especially and almost exclusively corpora of news sources reporting the same events. Barzilay and Lee[12] generated paraphrase sentences from news articles using finite state automata. Shinyama and Sekine[13] extracted paraphrases through the detection of named entities anchors in a corpus of Japanese news articles. Also working on Japanese news articles, Fujita and Inui[14] developed morpho-syntactic paraphrasing patterns to collect paraphrases.

We aim at detecting paraphrases in the same line as Elhadad and Sutaria[10] but for French. For this reason, we do not have available corpora of parallel lay and specialized documents. Besides, it is more difficult to obtain comparable corpora with a high level of similarity, such as news articles reporting the same events used in general language[12,13,14]. This led us to turn to an alternative and potentially more general approach, that is to use non-parallel, comparable corpora with less similar but more easily available documents—documents dealing with a same wide-range medical topic (such as nicotine addiction) but from different genres (lay and specialized). We describe our experiment in building and exploiting this type of corpora to find paraphrases between specialized and lay language. Issues at stake involve: (i) how to collect corpora as relevant as possible; (ii) how to identify passages which potentially convey comparable information; and (iii) what sorts of paraphrases can be collected between the two varieties of language. A common hypothesis is that specialized language uses more nominal constructions where lay language uses more verbs instead. We test this hypothesis and build on it to detect specialized-lay paraphrases around noun-to-verb mappings. We summarize the method we used to address points (i) and (ii)[15] and address point (iii) through the identification of nominalization paraphrases.

Material and Methods

Acquisition of comparable corpora of specialized and lay texts

For this work we acquired and used two different comparable corpora of specialized and lay texts: one corpus (the development corpus) which served as base material for the design and implementation of our paraphrase acquisition method; and a second corpus (the test corpus) on which we tested the method. We chose to work with medical documents dealing with the topic of diabetes for our development corpus and with the topic of tobacco and nicotine addiction for our test corpus. We thus built two comparable French corpora containing lay texts and specialized texts.

We collected our corpora from the Web, a popular source of corpora as it provides easy access to a virtually unlimited number of documents. When dealing with a Web corpus several issues arise including relevance to the targeted domain and to the targeted genre. We discarded generic search engines since the documents would not be categorised and relevance to the domain would be more questionable and decided in favour of a more restricted search involving a small part of manual work. We followed the same process for both corpora. First we queried two health search engines that we knew of (the health web portals CISMeF and HON[⋆]) with a list of keywords. They provide access to trustworthy Web pages and allow the user to search for documents targeted to a population. We also knew of relevant websites. Those were French governmental websites issuing documents for professionals and for lay people (that of the HAS and the INPES[†]); as well as health websites dedicated to lay people, including Doctissimo, Tabac Info Service and Stop-tabac[‡]. Once collected, a corpus needs to be cleaned and converted into an appropriate format, *i.e.* extracting the textual content. We converted our documents to text using methods described in previous work[15]. As part-of-speech tagging and lemmatization are needed in subsequent tasks, we tagged and lemmatized the corpus using the TreeTagger part-of-speech tagger[⋆⋆] and the French lemmatizer Flemm[⋆⋆⋆§].

Aligning lay and specialized passages

As a first step, we tried to relate text passages taken from both sides of our comparable corpora which address similar topics and might thus contain paraphrases. We proceeded in three steps: (1) as multiple topics are usually addressed in a single text, we performed topic segmentation on each text using the TextTiling[16] segmentation tool. A segment may correspond to one or several paragraphs; (2) we then tried to identify pairs of text segments addressing similar topics and likely to contain paraphrases. For this we used a common, vector-based measure of text similarity: the Cosine similarity measure which we computed for each pair of topic segments in the cross-product of both corpus sides (each segment was represented as a bag of words); (3) we selected the best text segment pairs, that is the pairs with a similarity score equal or superior to 0.33, a threshold we determined based on the results of a preliminary study[15].

Paraphrase acquisition between lay and specialized text segments

From the paired text segments we sought to detect paraphrases of specialized and lay language.

Preliminary study. We are looking for paraphrases between two varieties of language (specialized and lay), as opposed to any kind of possible paraphrases. We therefore need to determine what kind of paraphrases may be relevant in this regard. As a preliminary step to our study we computed statistics on our development corpus (the diabetes corpus), involving mainly frequencies of words and parts-of-speech. One of the findings of this step was that nouns tended to be slightly more numerous in the specialized texts while verbs tended to be slightly more frequent in the lay texts of our corpus. We also examined the pairs of topic segments from this corpus to try to determine what sort of paraphrases might be relevant. We noticed that in several segment pairs we had a deverbal noun

[⋆] http://www.cismef.org/ and http://www.hon.ch/

[†] http://www.has-sante.fr and http://www.inpes.sante.fr/

[‡] http://www.doctissimo.fr/ and http://www.tabac-info-service.fr/ and http://www.stop-tabac.ch/

[⋆⋆] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[⋆⋆⋆§] http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.html (e.g., "treatment") in the specialized segments while we had a verb (e.g., "treat") in the corresponding lay segments. So we hypothesized that one trend of specialized texts as opposed to lay texts might be to use deverbal nouns where lay texts more often use corresponding verb constructions. This hypothesis was supported by the corpus statistics. We therefore tried to detect paraphrases corresponding to that pattern.

Design of paraphrasing patterns. We used a lexicon of French deverbal nouns paired with corresponding verbs[17] to detect such pairs in the corpus segments. These pairs served as anchors for the detection of paraphrases. In order to design paraphrasing patterns we extracted all pairs of deverbal noun and verb with their contexts from the development corpus. The study of such pairs with their contexts allowed us to establish a set of lexico-syntactic paraphrasing patterns. An example of such patterns can be seen in table 1. The general method was to look for corresponding content words (mainly noun and adjective) in the contexts. We defined corresponding words as either equals or synonyms (we used lexicons of synonyms as resources[¶]). Equals have either the same part-of-speech, or different parts-of speech, in which case stemming is performed to take care of derivational variation (*e.g.*, "medicine" and "medical"). We then applied the patterns to both development and test corpora.

Table 1: Example paraphrasing patterns (a shared index indicates equality or synonymy. N=noun, V=verb, A=adjective, PREP=preposition, DET=determiner, 1 in index = pair of deverbal noun and verb)

Specialized	Lay
N_1 PREP (DET) N_2	V_1 (DET) N_2
N_1 PREP (DET) N_2A_3	V_1(DET) N_2A_3
N_1 A_2	V_1(DET) N_2

Evaluation

We evaluated the following points:

1. the quality of the results, *i.e.* if the detected paraphrases are indeed correct. For this we used a standard evaluation measure in NLP, precision, which is the percentage of correct results over the whole results. We also investigated the influence of using synonyms in the paraphrasing patterns, whether they brought too many incorrect paraphrases and whether they allowed to detect more paraphrases. So we ran two different implementations, one using the synonyms (syn) and one without using them (nosyn);

2. the quantity, *i.e.* whether we retrieved enough paraphrases. We investigated this point in two ways. First, we measured recall (that is the percentage of correct extracted paraphrases over the total number of paraphrases that should have been extracted). For this we used a random sample of 10 segment pairs from our test corpus from which we manually extracted paraphrases so as to set up a gold standard. Second, we also looked at the number of different deverbal noun/verb pairs involved in the paraphrases. We measured this number in the paraphrases found between the paired segments, but also in the paraphrases that can be found between any of the segments (*i.e.* in the set of segments involved in pairs as a whole) to see whether many types of noun/verb pairs were missed by limiting ourselves to the pairs. Additionally we measured the precision of paraphrases found between any of the segments to see whether the quality was any different, in case the quantity was much higher and might cause us to question the relevance of using segment pairs.

3. the coherence of the results with our hypothesis that deverbal nouns are more often used in specialized texts while lay texts tend to replace them with verbs. For this we computed the conditional probability $P(V_l|N_s)$ that a deverbal noun N_s in a specialized segment be replaced by a corresponding verb V_l in a corresponding lay segment. It can be estimated by $\frac{f(N_sV_l)}{f(N_s)}$, where $f(N_sV_l)$ is the number of occurrences of deverbal nouns in specialized segments having corresponding verbs in corresponding lay segments (*i.e.* the total number of deverbal noun/verb pairs that can be found between the specialized and lay paired segments) and $f(N_s)$ is the number of occurrences of deverbal nouns in the specialized segments.
To test whether this tendency of using verbs instead of nouns is indeed stronger in lay texts we also measured the reverse, *i.e.* the conditional probability $P(V_s|N_l)$, given a deverbal noun N_l in a lay segment, that it be replaced with a verb V_s in the corresponding specialized segment, computed as $\frac{f(N_lV_s)}{f(N_l)}$. If our hypothesis is verified the probability should be lower.

Results

Figures for the acquisition of the corpora and the alignment of text segments are given in table 2.

Quality

Evaluation of the quality of the detected paraphrases shows that precision is good for both implementations and for both corpora (see table 3), though higher for the implementation not using synonyms (nosyn); but as can be expected, paraphrases are less numerous in

[¶] The lexicons used came from the Masson and Robert dictionaries.

Table 2: Corpus sizes and number of segment pairs ("docs" = documents, "spec." = specialized, "dev." = development, "seg." = segment)

	Dev. corpus		Test corpus	
	Spec.	*Lay*	*Spec.*	*Lay*
docs	135	600	62	620
words	580,712	461,066	595,733	603,257
seg. pairs	183		547	

that case. Examples of paraphrases (and their English translations) are given in table 4. The last line shows an example of incorrect paraphrase.

Table 3: Number of different paraphrases (pa.) and precision with (syn) and without synonyms (nosyn)

	Dev. corpus		Test corpus	
	syn	*nosyn*	*syn*	*nosyn*
different par.	42	26	79	52
correct par.	30	22	62	49
precision	71.4%	84.6%	78.5%	94.2%

Table 4: Examples of detected paraphrases

Specialized	Lay
consommation régulière *regular use*	consommer de façon régulière *to use in a regular fashion*
gêne à la lecture *reading difficulty*	empêche de lire *prevents from reading*
évolution de l'affection *evolution of the condition*	la maladie évolue *the disease is evolving*
prise en charge *the taking care of*	prendre du poids *to put on weight*

Quantity

With regard to the quantity evaluation, we measured a 30% recall on our sample of segment pairs, meaning that out of the 10 manually extracted paraphrases only 3 were automatically detected by our method. Cases of non-detected paraphrases were due to the restrained scope of the paraphrasing patterns, as well as to the presence of synonyms not contained in our lists. Table 5 gives the number of different deverbal noun/verb pairs in the paraphrases found in paired segments compared to those found in the set of segment pairs as a whole (*i.e.* not necessarily belonging to the same pairs). Their number is higher in the whole set of segments than those retrieved in segment pairs, thus indicating that the step of aligning text segments causes the method to miss some paraphrases. However precision is lower for those paraphrases (last line of table 5) than for those found in segment pairs (table 3).

Table 5: Number of different noun-verb pairs in paraphrases (N-V paraph.) found in paired segments (*PS*) and those found in the set of segment pairs as a whole (*S*), and precision of paraphrases in *S* (using *syn*)

	Dev. corpus	Test corpus
N-V paraph. in *PS*	21	28
N-V paraph. in *S*	65	82
precision for paraph. in *S*	66.7%	59.3%

Coherence with the initial hypothesis

Table 6 displays results for the investigation on the coherence of our initial hypothesis that deverbal nouns in specialized texts tend to be replaced by verbs in lay texts. Deverbal nouns are more frequent in specialized texts, as are those having corresponding verbs in lay texts (noun/verb pairs). The conditional probability is also higher for specialized nouns paired with lay verbs than for the reverse order, which means that deverbal nouns in specialized texts are more likely to be replaced by verbs in lay texts than deverbal nouns in lay texts by verbs in specialized texts.

Table 6: Number of deverbal nouns and noun/verb pairs in the paired segments in both orders, specialized-lay (s-l) and lay-specialized (l-s)

	Dev. corpus		Test corpus	
	s-l	*l-s*	*s-l*	*l-s*
Nouns $f(N_s)$, $f(N_l)$	2,353	1,674	14,685	9,072
Noun-verb pairs $f(N_sV_l)$, $f(N_lV_s)$	1,213	602	5,142	2,505
Cond. prob. $P(V_l \vert N_s)$, $P(V_s \vert N_l)$	0.52	0.36	0.35	0.28

Discussion

In this work we built comparable corpora of specialized and lay texts on which we implemented a paraphrase acquisition method to extract a certain type of paraphrases that seemed relevant in the context of specialized and lay language, *i.e.* paraphrases based on deverbal noun vs. verb constructions. The precision measured on the set of detected paraphrases is rather high, which indicates good quality of the paraphrases (hence of the patterns and extracted segments). Although the quality decreases a little when using broader paraphrasing patterns involving synonym recognition, we think that this type of pattern is more desirable since it brings more paraphrases and also paraphrases which are more interesting since less direct in their identification. The originality of this work lies in the fact that, as opposed to approaches working with more closely related corpora (parallel[11,10] or very similar comparable corpora[12,14,13]), we gathered comparable corpora of documents which, although addressing the same general topics (nicotine addiction, diabetes), were a priori rather different since coming from various sources and targeted to different populations. We showed that simple paraphrase acquisition methods could also work on documents with a lesser degree of similarity, once similar segments were detected.

The quantity of extracted paraphrases is one drawback of this work. Recall is low which is mainly due to the fact that we set up rather rectricted paraphrasing patterns. This was done to ensure a high precision but caused the recall to fall. A future step would be to improve recall by modifying some aspects of the paraphrasing patterns while trying to keep a good precision. Regardless of recall, the number of paraphrases in itself is also small which can be due to the fact that we restrict ourselves to one specific type of paraphrases, but also to the facts that we first align and select similar text segments, that the coverage of our corpora might not be sufficient, and that we work on comparable corpora of lesser similarity than other methods. Future work to increase the number of paraphrases involves using clusters of text segments instead of pairs, increasing the corpus sizes and developing methods to detect other types of paraphrases.

Being in the context of specialized vs. lay language we had to spot types of paraphrases that might be relevant in this regard. We based our work on the hypothesis that among relevant types were paraphrases involving deverbal noun vs. verb contructions, meaning that lay texts tend to use verb contructions where specialized texts use deverbal noun contructions. The hypothesis was supported by our results. Such paraphrases therefore seem to be interesting advice to give to authors of lay texts. Future work includes testing our method on English and comparing the results for the two languages. We would expect them to be fairly similar since the tendency to use nominal constructions in scientific literature has also been observed for English[18].

Conclusion

To sum up, we described a method to identify paraphrases of deverbal noun and verb constructions in corpora of lay and specialized medical French texts, as a first step towards bridging the gap between these two varieties of language. We were able to identify paraphrases with a rather high degree of precision that seemed to reflect some of the systematic differences between specialized and lay texts. This is encouraging evidence since we worked with comparable corpora with a small degree of similarity. The next stage of our research is to contribute to authoring patient-oriented documents by proposing lay constructions that are more familiar to patients as a replacement of specialized constructions.

References

1. Lerner E, Jehle D, Janicke D, and Moscati R. Medical communication: Do our patients understand? *Am J Emerg Med* Nov 2000;18(7):764–6.
2. McCray A. Promoting health literacy. *J Am Med Infom Assoc* Mar-Apr 2005;12(2):152–63.
3. Leroy G, Eryilmaz E, and Laroya BT. Health information text characteristics. In: AMIA, (vol11), 2006:479–83.
4. Grabar N, Krivine S, and Jaulent MC. Cross-language classification of health web pages as expert and non expert. In: AMIA Annual Symposium, Chicago. 2007.
5. McCray AT, Ide NC, Loane RR, and Tse T. Strategies for supporting consumer health information seeking. In: Medinfo, (vol11), 2004:1152–6.
6. Zielstorff R. Controlled vocabularies for consumer health. *Biomed Inform* Aug-Oct 2003;36(4-5):326–3.
7. Zeng Q and Tse T. Exploring and developing consumer health vocabularies. *J Am Med Infom Assoc* Jan-Feb 2006;13(1):24–9.
8. Elhadad N. Comprehending technical texts: Predicting and defining unfamiliar terms. In: AMIA, 2006:239–43.
9. Zeng Q, Goryachev S, Kim H, et al. Making texts in electronic health records comprehensible to consumers: A prototype translator. In: AMIA, 2007.
10. Elhadad N and Sutaria K. Mining a lexicon technical terms and lay equivalents. In: ACL BioNLP Workshop, Prague, Czech Republic. 2007:49–56.
11. Barzilay R and McKeown K. Extracting paraphrases from a parallel corpus. In: ACL/EACL, 2001:50–7.
12. Barzilay R and Lee L. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In: HLT-NAACL, 2003:16–23.
13. Shinyama Y and Sekine S. Paraphrase acquisition for information extraction. In: International Workshop on Paraphrasing (IWP), 2003.
14. Fujita A and Inui K. A class-oriented approach to building a paraphrase corpus. In: International Workshop on Paraphrasing (IWP), 2005:25–32.
15. Deléger L and Zweigenbaum P. Aligning lay and specialized passages in comparable medical corpora. *Stud Health Technol Inform* 2008;136:89–94.
16. Hearst M. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 1997;23(1):33–64.
17. Hathout N, Namer F, and Dal G. An Experimental Constructional Database: The MorTAL Project. In: *Many Morphologies*. 2002:178–209.
18. Fang Z. Scientific literacy: A systemic functional linguistics perspective. *Science Education* 2005;89(2):335–47.